COMBINA-PT: A Large Corpus-extracted and Hand-checked Lexical Database of Portuguese Multiword Expressions

Amália Mendes, Sandra Antunes, Maria Fernanda Bacelar do Nascimento, João Miguel Casteleiro, Luísa Pereira and Tiago Sá

Centro de Linguística da Universidade de Lisboa (CLUL)

Av. Prof. Gama Pinto, 2, 1649-003 Lisboa, Portugal

{amalia.mendes, sandra.antunes, fbacelar.nascimento, luisa.alice, ptsa}@clul.ul.pt, miguel.casteleiro@zmail.pt

Abstract

This paper presents the COMBINA-PT project, a study of corpus-extracted Portuguese Multiword (MW) expressions. The objective of this on-going project is to compile a large lexical database of multiword (MW) units of the Portuguese language, automatically extracted from a balanced 50 million word corpus, interpreted with lexical association measures and manually validated. MW expressions considered in the database include named entities and lexical associations with different degrees of cohesion, ranging from frozen groups, which undergo little or no variation, to lexical collocations composed of words that tend to occur together and that constitute syntactic dependencies, although with a low degree of fixedness. This new resource has a two-fold objective: (i) to be an important research tool which supports the development of MW expressions typologies and their lexicographic treatment; (ii) to be of major help in developing and evaluating language processing tools able of dealing with MW expressions.

1. Introduction

Word Combinations in Portuguese Language (COMBINA-PT) is an ongoing project consisting of a large lexical database of Portuguese multiword (MW) expressions automatically extracted through the analysis of a balanced 50 million word corpus, statistically interpreted with lexical association measures and validated by hand.

The availability of large amounts of textual data and corpus-driven analysis enable the identification and analysis of complex patterns of word associations, showing that the lexicon is populated with chunks, some frozen, others not totally fixed but more or less predictable (Firth, 1955; Sinclair, 1991). These word associations, when frequently repeated, tend to correspond to a conventional way of saying things, turning out to be an important aspect in the lexical structure of the language.

The COMBINA-PT database covers word associations with different degrees of cohesion, ranging from totally frozen groups, semi-frozen or just sets of favoured co-occurring forms, as well as named entities.

The vast corpus used and the powerful automatic processing tools devised assure a large coverage of Portuguese word associations that is of major importance for the main objectives of this new resource: to be an important research tool which supports the development of collocation typologies and their integration in a larger theory of MW units and to be of major help in developing and evaluating language processing tools able of dealing with MW units.

This paper will discuss the corpus constitution and the MW unit's extraction tool in section 2, the lexical database implementation and the methodology adopted in section 3 and further developments of the work undertaken in section 4.

2. Corpus and Multiword Unit's Extraction Tool

The COMBINA-PT corpus is a balanced 50,8M word written corpus that was designed and compiled using resources available from the Reference Corpus of Contemporary Portuguese (CRPC). CRPC is a written and spoken monitor corpus (cf. Sinclair, 1991), compiled at CLUL since 1988 and that comprises all Portuguese geographical varieties, in a total of 330 million words (http://www.clul.ul.pt/english/sectores/projecto_crpc.html).

The corpus balance is essential to assure that specific textual and discursive patterns of co-occurrence of a lexical unit are uncovered, given that a particular word may co-occur with different lexical units according to the type of discourse in which it occurs.

According to this criterion, the corpus design covers different genres presented in table 1, below:

CORPUS CONSTITUTION			
Newspapers			**30.000.000**
Books	Fiction	6.237.551	
	Technical	3.827.551	
	Didactic	852.787	**10.818.719**
Magazines	Informative	5.709.061	
	Technical	1.790.939	**7.500.000**
Miscellaneous			**1.851.828**
Leaflets			**104.889**
Supreme court verdicts			**313.962**
Parliament sessions			**277.586**
TOTAL			**50.866.984**

Table 1: Constitution of the Corpus

An interesting development of our study will be to run the MW extraction software on the Portuguese

spoken corpus of 1M words, previously compiled at CLUL. The discrepancy between the available amount of written and spoken corpus makes it necessary to process the data separately in a latter stage of the project.

The MW unit's extraction tool automatically extracts from the corpus groups of 2, 3, 4 and 5 tokens (groups of 2 tokens can be contiguous or be separated by a maximum of 3 tokens, while groups of more than 2 tokens are contiguous) and gives several types of information regarding each group, as showed in Table 2 with the example *fio de prumo* 'plumb line'.

Results identify the group of words (in bold), the distance between the group word forms (first number after the MW unit in bold) and how many elements form the chunk (eg). The tool counts the number of occurrences of the group in the corpus at the specific distance mentioned (og), the total number of occurrences of the group at any distance (fg), the number of occurrences of each element of the group (fe) and the total number of tokens in the corpus (N). This frequency information is used to statistically analyse chunks extracted by applying a selected association measure and sorting the results.

The tool allows the user to select which measure to apply, and was first run with Mutual Information (MI) that calculates the frequency of each group in the corpus and crosses this frequency with the isolated frequency of each word of the group, also in the corpus (Church & Hanks, 1990). The results obtained when running the tool on the 50M tokens corpus range from a minimum MI value of -5 to a maximum of 33 and are expressed in Table 2 as the (ic) value. Finally, the tool also extracts the concordance lines of the MW expression and presents them in KWIC format with a reference code. The size of the context window can be established when running the extraction tool, as well as the sorting option.

6 **fio de prumo** 1 eg(3) og(6) ic(9.844055) fg(6) fe(1877 2290575 71) N(50310890)

123962906	indicada, para cada ponto, pelo	fio de prumo;	- o @bsentido@b-
123962913	erces e alinhar as paredes com o	fio de prumo.	A casa gandaresa é
123962920	s bastavam para saber utilizar o	fio de prumo	e travar bem os ado
123962927	á o músico António Pinho Vargas,	fio de prumo	(móveis, design, ex
123962934	prumada do edifício, tendo-se o	fio de prumo	prendido num grampo
123962941	nosso comandante!: recto como um	fio de prumo.	Rico homem!... ALB

Table 2: Example of the MW unit *fio de prumo* 'plumb line'

Several cut-off options are also available in order to reduce noise: (i) groups with internal punctuation can be eliminated since MW expressions show at least some degree of cohesion and do not allow strong punctuation; (ii) two-word groups with initial or ending grammatical words can also be eliminated using a stop-list if one wants to study lexical associations instead of functional compounds (e.g., compound prepositions) or verb valency; (iii) a minimum frequency can also be selected. When running the tool on the corpus these 3 options were selected, with a minimum frequency of 3 to groups of 3 to 5 tokens and of 10 for 2-token groups. Despite the implementation of these cut-off options, the results obtained were still considerable since the candidate list comprises 1,7M units.

Neither the corpus nor the extracted MW expressions were tagged. Although it is true that POS information could be useful to select syntactic patterns and make it easier to recover significant units, it would also lead us to search for specific well-known patterns, while we were aiming at discovering those patterns with a corpus-driven approach that would further legitimate a typology of MW expressions. Besides, neither manual revision nor lemmatization of the 50M word corpus could have been achieved, making the results less trustful.

3. Database of Lexical Collocations

The results attained were then loaded into a MySQL database, a relational database with a client/server typology, designed to enable the representation of the MW units and to offer a platform for user-friendly manual validation. The database includes all the fields exemplified in Table 2 and is directly associated to the corpus text and the index file so as to allow the user to view concordance lines in a wider window context, since manual validation is strongly dependent on the observation of the context. Concordance lines are sometimes wrongly associated with a MW unit (for example, when the same sequence of words can function as a MW unit with a non-compositional meaning or as a totally free combination with compositional meaning) and must be eliminated during the hand-validation process. In those cases, the total group frequency is automatically recounted in the Frequency field.

Since the exact definition of a collocation and how it differs from other MW expressions is known as a challenging issue (discrete categorization is difficult to apply to concepts defined in terms of degree of fixedness, compositionality, substitutability, etc.) and considering the large set of groups to be covered, it was decided that, at a first stage of the work, we would select all the expressions that presented some syntactic and semantic cohesion, without attempting to follow any prior typology. We did, however, try to organize our data according to their function and internal structure, in order to identify MW expressions that:

- may or may not occur with an hyphen (e.g., *pronto a vestir* 'ready to wear'; *caminhos de ferro* 'railway');
- refer to named entities, such as institutions, functions, etc. (e.g., *Parlamento Europeu* 'European Parliament', *Escola de Belas-Artes* 'Fine Arts School');
- constitute verbal phrases (e.g., *respirar fundo* 'to breathe deeply');
- constitute other phrases, like nominal phrases (e.g. *condições de trabalho* 'work conditions') or adjectival phrases (e.g. *meramente formal* 'merely formal');
- are doubtful cases or that exceed the maximum number of 5 tokens extracted by the tool (these cases will have to be correctly identified during the lemmatization process).

A list of all the word forms present in the selected MW units is automatically created. Manual validation can also be processed through the list of all inflected forms in the candidate list, since each inflected form is associated with a list of all MW expression it enters in. An example of a record of the database is presented in figure 1.

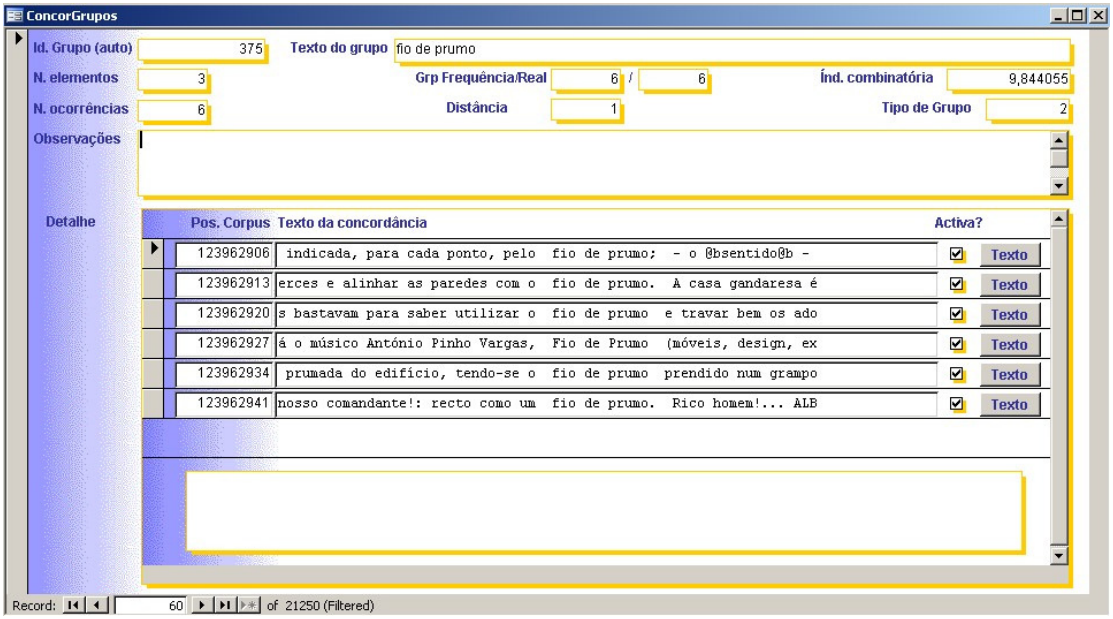

Figure 1: Record for the collocation *fio de prumo* 'plumb line' in the database

Although no prior typology was followed during the manual validation process, a first analysis of the expressions selected point to different types of MW units, with different degrees of cohesion, ranging from:

— fully lexicalized groups which show no (or minimum) variation such as proverbs or idioms and which do not undergo inflectional variation nor internal modification (e.g., grão a grão enche a galinha o papo 'many a mickle makes a muckle');

— not fully lexicalized groups with non-compositional meaning (e.g., *fazer ouvidos de mercador* 'to turn a deaf ear'). These expressions are not subject to internal modification (**fazer muitos ouvidos de mercador* 'to turn a very deaf ear') nor to passivization (**ouvidos de mercador foram feitos* 'deaf ear was turned'), but they can undergo inflectional variation, especially when one of the elements is a verb (e.g., *fizeram ouvidos de mercador* '[they] turned a deaf ear');

— not fully lexicalized groups that can have either compositional or non-compositional meaning and that allow for the substitution of one of the collocates by other words of the same semantic domain (onda/maré/vaga de assaltos 'wave/tide of robberies');

— groups that are fully compositional but that are, however, favoured co-occurring forms since they occur with much higher frequency than any other alternative lexicalization of the same concept, which reveals that they may be in their way to a possible fixedness (rajada de vento 'blast of wind'; físico e psicológico 'physical and psychological').

This first try at a typology of MW expressions shows that even lexicalized groups can undergo some inflectional variation. There is of course a large set of MW expressions that only occur in the corpus in a specific word form, like the case of the nominal phrase reparação de danos 'damage repair', that never occur in the plural form *reparações de danos 'damage repairs'.

But, being Portuguese a highly inflectional language, like other romance languages, most MW expressions do accept inflectional variation on one or even all of the group elements, making it necessary to lemmatize the set of MW expressions and to associate groups with each lemma. The fact that prepositions can contract with the following article/pronoun creates an even greater word form variation. For example, the group dar uma espreitadela a 'take a peep at' can present verb inflectional variation as well as different word forms for different contractions of preposition a 'to' and the following article or pronoun (e.g., deu uma espreitadela à '[he/she] gave a peep at_the[fem, sg]', deram uma espreitadela aos '[they] gave a peep at_the[masc, pl]', dei uma espreitadela àquela '[I] gave a peep at_that_one[fem, sg]' - contracted elements are connected in the English translation).

When a MW expression is spread into several inflectional variants, it is possible that none of its variants reaches the minimum frequency to make the group automatically recognized as a possible significant expression. This is true for many expressions with a verbal element, like *esfregar as mãos de contentamento* 'to rub ones hands with satisfaction' where the different word forms of the verb *esfregar* 'rub' have very low frequency (frequency 1, 2 or 3 maximum). Since a minimum frequency was established during the tool running process, none of the group realization is recovered and it is the visualization of the concordance lines of a smaller group *as mãos de contentamento* that points to the existence of a larger expression (see figure 2).

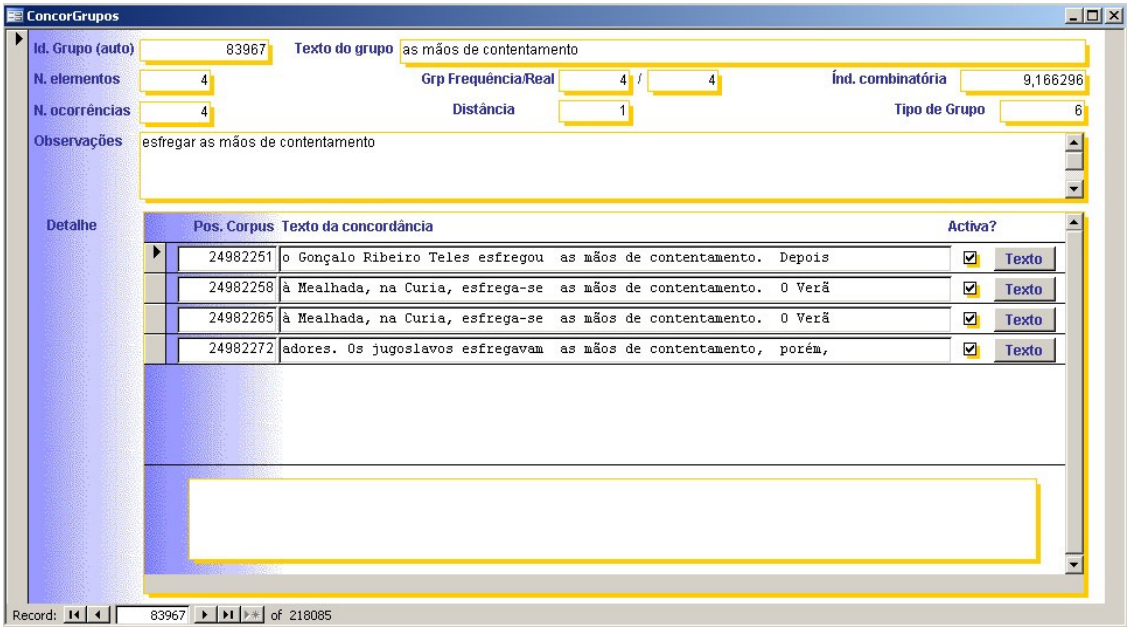

Figure 2: Record for the group *as mãos de contentamento* 'hands with satisfaction'in the database

Considering the large candidate list of 1,7M units, it was impossible to assure manual validation of the whole data, which led us to hand-check only a subpart of the list. The selection of this subpart relied on the information given by the MI association measure. Previous work on MW expressions in Portuguese (Bacelar do Nascimento, 2000; Pereira & Mendes, 2002) as well as manual analysis of specific lemma during this project showed that MI gave better results with medium values between 7-11, a conclusion similar to those of evaluative studies of association measures like Evert & Krenn (2001). For example, the MW expression *fio de prumo* 'plumb line', clearly lexicalized, receives a medium MI value of 9,8. We thus started to manually validate MW units with MI values between 8 and 10, a total of 170,000 units, 10% of the total candidate list. Of those, 31,000 were selected as significant expressions and 1,637 were considered doubtful and will be revised.

From these 31,000 MW expressions already selected, a list of lemma that compose these expressions is automatically created so that the hand-checking process will continue with the validation of all the remaining MW units of the total candidate list that comprise those lemma.

4. Further Developments

The data collected and its analysis will make it possible to propose a corpus-driven typology of MW expressions.

Besides the issues on fixedness degree and compositional meaning, the study of these MW expressions allows to identify associative patterns that characterizes a word according to: (i) co-occurrence patterns (systematic co-occurrence with particular lexical items in a contiguous or non-contiguous form); (ii) grammatical patterns (systematic co-occurrence with a certain verb class, with specific temporal verb forms or with certain syntactic constructions); (iii) paradigmatic

patterns (hyperonymy, homonymy, synonymy or antonymy phenomena); (iv) discursive patterns (strong associations in one language register can be a weak association in another register).

The database will enable a systematic analysis of the structure of the lexicon, by giving us information on the number of chunks that compose the lexicon (either compound nouns, lexical collocations or totally lexicalized expressions) and by allowing a quantitative comparison between free word forms and MW expressions.

This lexical database will be of extreme importance for several areas, ranging from psycholinguistics (development of hypothesis about the representation of the individual mental lexicon, semantic memory and cognitive processes), lexicography (improvement of their coverage in modern dictionaries) or computational linguistics (helping to develop and evaluate language processing tools able of dealing with MW expressions specific issues, like automatic unit recognition, lexical association measures for validation of significant MW units, as well as overgeneration, tagging and parsing problems (Sag *et alii*, 2002)).

The Lexical Database of hand-checked MW units will be available for online query at the project site: http://www.clul.ul.pt/english/sectores/projecto_combina.html.

6. Acknowledgments

The Word Combinations in Portuguese Language (COMBINA-PT) project is undertaken at the Centre of Linguistics of the University of Lisbon under a research grant of the Portuguese Ministry of Science (POCTI/LIN/48465/2002).

7. References

Bacelar do Nascimento, M. F. (2000) "Exemples de combinatoires lexicales établis pour l'écrit et l'oral à Lisbonne", in Bilger, M. (ed.) *Corpus, Méthodologie et Applications Linguistiques*, Paris, H. Champion et Presses Universitaires de Perpignan, pp. 237-261.

Bahns, J. (1993) "Lexical collocations: a contrastive view", *ELT Journal*, 47:1, pp. 56-63.

Biber, Douglas (1996) "Investigating language use through corpus-based analyses of association patterns", *International Journal of Corpus Linguistics*, vol 1 (2), pp. 171-197.

Braasch, A. & S. Olsen (2000) "Toward a Strategy for a Representation of Collocations – Extending the Danish PAROLE-lexicon", *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athens, Greece, 31 May – 2 June 2000, vol. II, pp. 1009-1016.

Butler, C. S. (1998) "Collocational Frameworks in Spanish", *International Journal of Corpus Linguistics*, vol. 3(1), pp. 1-32.

Calzolari, N. et al. (2002) "Towards Best Practice for Multiword Expressions in Computational Lexicons", *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas, Canary Islands; Spain, 29 May – 31 May 2002, pp. 1934-1940.

Church, K. W. & P. Hanks (1990) "Word association norms, mutual information, and lexicography", *Computational Linguistics*, 16 (1), pp. 22-29.

Clear, J. (1993) "From Firth principles: Computational tools for the study of collocation", in Baker, M., G. Francis and E. Tognini-Bonelli (eds.) *Text and technology: In honour of John Sinclair*, Amsterdam, John Benjamins.

Evert, S. & B. Krenn (2001) "Methods for the Qualitative Evaluation of Lexical Association Measures", *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France, pp. 188-195.

Firth, J. (1955) "Modes of meaning", *Papers in Linguistics 1934-1951*, London, Oxford University Press, pp. 190-215.

Firth, J. (1957) "A Synopsis of Linguistics Theory, 1930-1955", *Studies in Linguistic Analysis.* Oxford Philological Society; reprinted in Palmer, F. (ed.) (1988) *Selected Papers of J. R. Firth*, Harlow, Longman.

Hausmann, K. W. (1979) "Un dictionnaire des collocations est-il possible?", in *Travaux de Linguistique et de Littérature XVII*, 1.

Heid, U. (1998) "Towards a corpus-based dictionary of German noun-verb collocations", *Euralex 98 Proceedings*, Université de Liège, Belgique.

Kjellmer, G. A. (1994) *Dictionary of English Collocations*, Oxford, Oxford University Press.

Krenn, B. (2000a) *The usual suspects: Data-oriented models for identification and representation of lexical collocations*, German Research Center for Artificial Intelligence and Saarland University Dissertations in Computational Linguistics and Language Technology, vol. 7, Saarbrücken, Germany.

Krenn, B. (2000b) "CDB - A Database of Lexical Collocations", *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athens, Greece, 31 May – 2 June 2000, vol. II, pp. 1003-1008.

Krenn, B. (2000c) "Collocation Mining: Exploiting Corpora for Collocation Identification and Representation", *Proceedings of KONVENS 2000*, Ilmenau, Deutschland.

Mackin, R. (1978) "On collocations: Words shall be known by the company they keep", in *Honour of A. S. Hornby*, Oxford, Oxford University Press, pp. 149-165.

Mel'cuk, I. (1984) *Dictionnaire explicatif et combinatoire du français contemporain*, Les Presses de L'Université de Montréal, Montréal, Canada.

Mel'cuk, I. (1995) "Phrasemes in Language and Phraseology in Linguistics", in Everaert, M. et al. (eds.) *Idioms: Structural and Psychological Perspectives*, Hillsdale, NJ/Hove, UK, Lawrence Erlbaum Associates Publ., pp. 169-252.

Pearce, D. (2002) "A Comparative Evaluation of Collocation Extraction Techniques", *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Spain, pp. 13-18.

Pereira, L. A. S. & A. Mendes (2002) "An Electronic Dictionary of Collocations for European Portuguese: Methodology, Results and Applications", in Braasch, A. & C. Povlsen (eds.), *Preceedings of the 10th EURALEX International Congress*, Copenhagen, Denmark, vol. II, pp. 841-849.

Pereira, L. A. Santos (1994) *Como se combinam as palavras? Contributo para um Dicionário de Combinatórias do Português*, M.A. Thesis, Faculty of Letters, University of Lisbon, ms.

Sag, I., T. Baldwin, F. Bond, A. Copestake & D. Flickinger (2002) "Multiword Expressions: A Pain in the Neck for NLP", in Gelbukh, A. (ed.) *Proceedings of CICLing-2002*, Mexico City, Mexico.

Sinclair, J. & A. Renouf (1991) "Collocational Frameworks In English", in Aijmer, K. and B. Altenberg (eds.) *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, Longman, Harlow, pp. 128-143.

Sinclair, J. (1991) *Corpus, Concordance, Collocation*, Oxford, Oxford University Press.

Smadja, F. (1990) "Retrieving Collocations from Text: Xtract", *Computational Linguistics*, vol. 19:1, pp. 143-177.